DOI: http://dx.doi.org/10.26483/ijarcs.v9i2.5882

ISSN No. 0976-5697

Volume 9, No. 2, March-April 2018

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

GENDER ESTIMATION FROM FIRST NAME: A RULE BASED APPROACH

Monali Y. Khachane
Yashwantrao Chavan School of Rural Development,
Shivaji University, Kolhapur
Kolhapur, India

Abstract: This paper presents an approach to identify gender from the first name of users. The name identification is carried out by identifying similarities from male and female gender names. 300 male and female gender student names belonging to Hindu religion are chosen for the current research work.Decision Tree induction (J48) is used to categorized names into male and female class. The J48 produce 96% accuracy for 40 % training and 60% testing percentage split.

Keywords:Decision Tree Induction

1. INTRODUCTION

Indian Culture refers collectively thousands of distinct and unique cultures of all religions and communities in India. This makes India rich and make different from other countries. The Indian names are inspired by their culture and religion. So we found differences in names of male and female genders religion wise. In this paper we had undertaken only Hindu religion data for analysis purpose.

Recognition of genders and age from images using human face are commonly available features in today's Smart gadgets. Gender identification is used for user segmentation, targeting advertisements and gender wise research analysis. This can be done by using first names of users. User Images, Online profiles, User tweets are used by some of the researchers for gender identification.

Many of the Industries interested to find the gender wise target groups online. Until this data is used for different purposes like ad targeting, gender wise research. But use of this identification will be extended for other various purposes like we can use this data for targeting the gender wise groups for health awareness campaign for commonly found diseases in male and female groups,Women Empowerment and female gender will be targeted to aware about laws related to Womensafetyand Sexual harassment, also can be used to aware the government schemes for targeted groups.

2. LITERATURE REVIEW

Gallagher[1] combined image-based gender and age classifiers with the cultural context information provided by first names to recognize people. Researcher demonstrated a model for the relationship between first name, age, gender and appearance in images.

Sarthak Gupta [2] presented an approach to detect gender of a person through frontal facial image using Delaunay Triangulation with accuracy 93.8283%.

Wendy Liu [3] proposed a approach to predict Twitter user's gender from his/her online content. Researcher used user's self-reported first name for gender inference which shows 20% more accurate results than previous approaches. They generated a gender-labeled Twitter dataset without depending on textual content and made it available for the research community.

Emad AlSukhni[4] collected 4017 tweets of Arabic authors for gender prediction. Their results show that J48, KNN, Naïve Bayes, NBM and SVM provide more than 99% accuracy for gender prediction.They also reported preprocessing effects negatively on results. They used author names and word features for prediction.

KamilWais [5] recommended R package called as genderizeR to choose the optimal approach suitable for different gender analyses. The researcher reviewed the different data sources like US census, Quebec Cences, WikiNameportal, multiple Wikipedia pages, top-100-baby-names-search.com portal,US Social Security Administration records,Social Network profiles. The approaches used to assign gender by using suffix of a first name are Rule based approach and human coder.

When the literature is reviewed on this ara we found that Gender API is one of the online source for gender identification from first names. This online API produce results by comparing the string with database included 1,877,788 names from 178 countries with 99.9% availability. The portal assigns gender on the basis of available number of samples for that name. Which we found not suitable for Indian names because few names are used for male and female gender commonly.

3. METHODOLOGY AND RESULTS

Human first name is identity given to the human body by our parents. In this paper we analyzed the names of students belongs to Hindu religion to identify the patterns according to gender. The 300 student database. This unique names database including 137 girls and 163 boys is undertaken for this study. The Marathi language is written in devnagrilipi. Attempt is made to find out the similarities in each gender. Table 1 includes the commonly identified letter in each gender.

A. Findings:

[1] A girl names ends with last letter pronounces in devnagari as "आ", "इ" and "ई".In English last letters found in these names are "a", "e" and "i" all are vowels in English language.

[2] Boys names ends with last letter pronounced in devnagari as "अ" and "ऊ". It is noted that all letters found in these names are consonants of English language.

[3] Also the commonly identified devnagari letter is "ल".

[4] Very few names used commonly for boys and girls having last letter "ल" and "म". "न" is used in rare names.

[5] Boys names include 'a' when last letter are a joined devnagari letter. e.g. "अजिंक्य" , "Ajinkya"

Table 1: Commonly identified letters

Gender	*Last Letter*
female	a, e, i
Male	b,d,g,h,j,k,l,m,n,o,p,r,s,t,u,v,
Male/female	l, m

Fig.1 shows the Decision tree buildusing J48 Decision Tree Induction method.

These findings are used to automated identification of user's gender. Simple algorithm is written to identify gender from decision tree is as follows

1. Convert first name in English
2. Extract last letter from first name of user.
   a. If last letter is "a", "e" or "i" then gender= "female"

   Else if last letter is "l" or "m"
   Then
   Gender=male or female
   Else
   Gender = "male"
   End if

The proposed algorithm is successfully classified the name of students. The Decision Tree method is evaluated using different percentage split for training and testing. The comparative results are as shown in Table 2.

Table 2: Gender Estimation Result

Classifier	*Training*	*Testing*	*% accuracy*	*Total Names*
J48	50	50	95.77	284
	70	30	95.2	
	80	20	94.2	
	40	60	96.47	
OneR	50	50	95	
	40	60	95.24	

B. Applications

1) Targeting Gender groups:-
   These groups can be used for online advertisements, gender wise health awareness, gender wise social issue identification and solution finding, Group wise research analytics
2) User Segmentation
3) Automated gender selection in online form filling

4. CONCLUSION:

The proposed approach shows that gender can easily identified by extracting last letter of first name with 96% accuracy. One of the important limitations of the current work is only Hindu religion names are used for the categorization purpose. The proposed work will be extended in future to identify genders belongs to more religions.

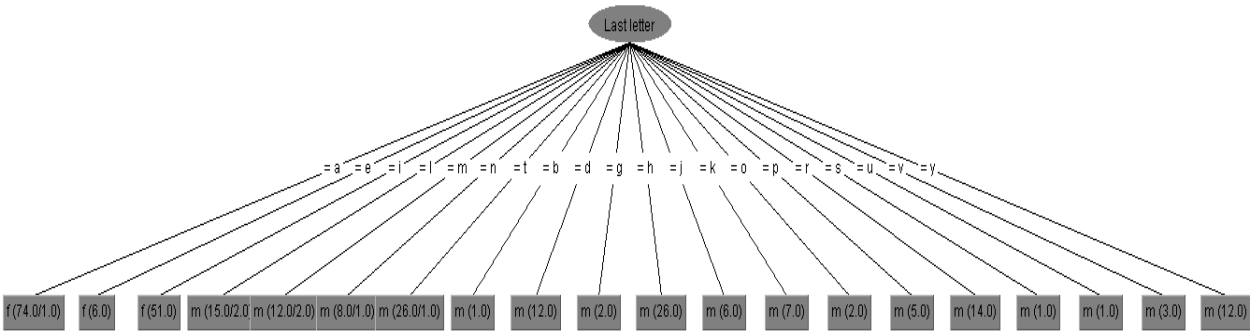

Figure 1: Decision Tree

REFERENCES

1. Gallagher, Andrew C., and Tsuhan Chen. "Estimating age, gender, and identity using first name priors." Computer Vision

and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.

2. Sarthak Gupta , "Gender Detection using Machine Learning Techniques and Delaunay Triangulation", International Journal of Computer Applications (0975 – 8887) Volume 124 – No.6, August 2015,pp. 27-32
3. Wendy Liu , Derek Ruths , "What's in a Name? Using First Names as Features for Gender Inference in Twitter", Analyzing Microtext: Papers from the 2013 AAAI Spring Symposium,pp.10-16
4. Emad AlSukhni , QasemAlequr , "Investigating the Use of Machine Learning Algorithms in Detecting Gender of the Arabic Tweet Author", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 7, 2016,pp.319-328
5. KamilWais , "Gender Prediction Methods Based onFirst Names with genderizeR", The R Journal Vol. 8/1, Aug. 2016 ISSN 2073-4859,pp. 17-37